Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-03-30 12:10:50
REFERENCE NO: 199

Author Names & Affiliations

- Ming Xian - Washington State University
- Shuiwang Ji - Washington State University

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Chemistry, Chemical Biology

Title of Submission

Data-driven Discovery of Chemical Tools: Cyberinfrastructure Needs and Student Training

Abstract (maximum ~200 words).

Chemical information is growing dramatically fueled by massive data generated from different researchers in the same or similar fields. This data has the potential to significantly advance science and technology in the corresponding fields. However, due to the lack of suitable databases and the appropriate computational approaches for data integration and analysis, data-driven development of small molecule chemical tools are often ignored or hard to pursue. To enable research ability in this area, advanced cyberinfrastructure (including advanced computing resources, data and software infrastructure, and student training programs) needs to be established at institutions like WSU. In this response to DCL, we describe 1) current challenges in the development of chemical tools for biomedical research; 2) the need for data-driven approaches for the development of next generation chemical tools; 3) cyberinfrastructure needed for chemical tool development and student training.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Chemical tools are widely used in biomedical research. They include a variety of small molecule compounds that can be used as enzyme inhibitors, drug donors, imaging sensors, specific protein tags, etc. The development of chemical tools is an ever expanding and fast growing field in chemical biology and a large number of chemical tools in each category have been reported. For example, small molecule fluorescent sensors are very useful in the study of sulfur-related redox biology. Hundreds of such sensors have been developed for specific sulfur-based biomolecules including glutathione, cysteine, homocysteine, hydrogen sulfide, etc. The availability and large scope of these sensors potentially provides sensor users with many options. On the other hand, such a large landscape of sensors can also confuse the users, as it is very difficult to decide which one is the most suitable sensor for a specific assay. The same problem also exists when

researchers are working on sensor development as they often struggle to keep up to date with the relevant literature. Because the data is large and unorganized, researchers are unwilling or unable to appropriately analyze known results. As such, many of them are doing somewhat redundant work. They usually mimic reported sensors and slightly modify the fluorophore or the reaction site of the known sensors to create 'new' ones. Although this type of research can produce structurally different sensors (sometimes with interesting properties), often times it ends up with marginal or no improvement from the known ones. It should be noted that the same problem exists for the development of many other chemical tools such as drug donors (such as hydrogen sulfide donors, nitric oxide donors), enzyme inhibitors, catalysts, etc. Therefore, it is a generic challenge in chemistry discovery.
To solve this problem, we believe the key is to have access to searchable and analyzable databases of known chemical tools. Taking fluorescent sensor as an example, ideally researchers would like to have access to a reliable and well-maintained sensor database, including their structures, photochemical/photophysical properties, application data, etc. Such a database will allow researchers to compare and select existing sensors for their studies. Moreover, combining the database with advanced data mining and machine learning techniques, we can significantly enhance our ability to create next generation sensors with improved properties.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

The storage, analysis, and sharing of growing chemistry data sets require software and hardware infrastructure at various stages. Most institutions provide computing support in terms of high-performance hardware and common software tools. However, we increasingly realize that these centrally-managed, common-purpose computing hardware and software are not enough to meet the need of chemistry data set analysis. For example, the analysis of chemical compound data requires the use of specialized software tools that are usually not available or not installed on common computing hardware. Usually, it is hard to request the installation of special purpose software tools on institutional hardware for security and other reasons. In addition, given the recent success in deep neural network based data analysis and machine learning methods, graphic processing units (GPUs) are increasingly used in processing structural data such as small molecule organic compounds. These high-performance hardware are not commonly available from institutional infrastructure. Thus, there is a need to provide both high-performance hardware and software specifically for big chemical data analysis. For example, the newly released NVIDIA DGX-1 AI Supercomputer (http://www.nvidia.com/object/deep-learning-system.html) is specifically designed for artificial intelligence and machine learning purposes such as deep learning technologies.
Data-driven chemistry discovery is a new research topic for many institutions like WSU. With appropriate cyberinfrastructure in place, this type of research will impact teaching and research significantly. It will facilitate and enhance the interaction between chemists and computer/data scientists. It will serve as a unique platform for training the next generation of scientists, especially chemists. Students can be involved in the projects at all stages: from database set up to data-driven chemical tool design; from organic synthesis to analytic chemistry evaluation. This set of skills is rarely taught in traditional hypothesis-driven chemistry projects but is becoming essential in this digital era. These students, able to work at the interface of chemistry, data resources, and advanced computing, are highly sought after by both industry and academia.

Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode)."